

AD _____

Award Number: **W81XWH-08-1-0731**

TITLE: **Grid-Enabled Quantitative Analysis of Breast Cancer**

PRINCIPAL INVESTIGATOR: **Andrew R. Jamieson, BA**

CONTRACTING ORGANIZATION:

**University of Chicago
Chicago, IL 60637**

REPORT DATE: **October 2009**

TYPE OF REPORT: **Annual**

PREPARED FOR: **U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012**

DISTRIBUTION STATEMENT:

☒ Approved for public release; distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YYYY) 01-10-2009		2. REPORT TYPE Annual Report		3. DATES COVERED (From - To) 1 Oct 2008 - 30 Sep 2009	
4. TITLE AND SUBTITLE Grid-Enabled Quantitative Analysis of Breast Cancer				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER W81XWH-08-1-0731	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) PI: Andrew R. Jamieson Collaborators: Maryellen L. Giger PhD, Karen Drukker PhD, Lorenzo Pesce PhD, Hui Li PhD, Neha Bhooshan, Yading Yuan				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of Chicago Chicago, IL 60637				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT The long-term goal of this research is to improve breast cancer diagnosis, risk assessment, response assessment, and patient care via the use of large-scale, multi-modality computerized image analysis. The central hypothesis of this research is that large-scale image analysis for breast cancer research will yield improved accuracy and reliability when optimized over multiple features and large multi-modality databases. We designed and executed a pilot study to utilize large scale parallel Grid computing to harness the nationwide cluster infrastructure for optimization of medical image analysis parameters. Additionally, we investigated the use of cutting edge data-analysis/mining techniques as applied to Ultrasound, FFDM, and DCE-MRI Breast Image Feature Space Analysis for CADx, specifically, dimension reduction and data representation techniques (t-SNE and Laplacian Eigenmaps) for high dimensional data spaces. These methods allow for an alternative to traditional feature selection methods. Using the 256-CPU high-throughput cluster computing capabilities, performance metrics and intensive statistical cross-validation (0.632+ bootstrap and ROC analysis for AUC performance) were performed to gain understanding of the new techniques potential versus previous Breast CADx methodologies. Results indicate the ability to rival or exceed previous state-of-the-art CADx performance.					
15. SUBJECT TERMS Multi-Modality Breast Image Analysis, Grid Computing, Dimension Reduction and Representation					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 51	19a. NAME OF RESPONSIBLE PERSON USAMRMC
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U			19b. TELEPHONE NUMBER (include area code)

Table of Contents

	<u>Page</u>
Introduction.....	4
Body.....	5
Key Research Accomplishments.....	9
Reportable Outcomes.....	10
Conclusion.....	11
References.....	12
Appendices Index.....	13
Appendix A.....	14
Appendix B.....	15
Appendix C.....	51

INTRODUCTION

Breast cancer is a leading cause of death in women, causing an estimated ~40,000 deaths per year. Mammography is the most effective method for the early detection of breast cancer, and it has been shown that periodic screening of asymptomatic women does reduce mortality[4]. Many breast cancers are detected and referred for surgical biopsy on the basis of a radiographically detected mass lesion or cluster of microcalcifications. Useful interpretation in mammography depends on the quality of the mammographic images and the ability of the radiologists who interpret those images. In addition to mammography, follow up imaging with the use of other modalities, such as MRI, and ultrasound are used for assessing malignancy of objects discovered following routine screenings. The long-term goal of this research is to improve breast cancer diagnosis, risk assessment, response assessment, and patient care via the use of large-scale, multi-modality computerized image analysis. The central hypothesis of this research is that large-scale image analysis for breast cancer research will yield improved accuracy and reliability when optimized over multiple features and large multi-modality databases. In recent years, data mining and data driven discovery have become important research tools in many disciplines. Massive amounts of data may contain hidden structure and rich information, previously unavailable for characterization within smaller subgroups. Systematic search can reveal that structure and information. In our context, the opportunity is as follows: The digital age of medical imaging provides an ever-growing archive of data. Deep analysis of this multimodal imaging data can be used to train and optimize algorithms that are incorporated into usable clinical systems, thus improving overall breast imaging interpretation and patient outcome. Data mining can also enable relational discoveries between image data and cancer diagnosis, response, and outcome, thus adding to the potential for “patient-specific diagnoses leading to patient-specific management.” Aspects of optimization in this process of CADx development, were previously infeasible due to massive data and computation requirements. However, now with advances in Grid-based computing many research avenues exist.

This reports covers the important initial developments in research accomplished in the past year leading towards these long term objectives. During this first year of research activity, in addition to executing a proof of principle Grid-based computing work-flow, the recipient was mainly focused on identifying suitable theoretical/analytical tools for carrying out larger scale investigation as described above, and conducting preliminary evaluations of the usefulness of these new methods.

BODY

Training Accomplishments

At the time of this report, the recipient of the Predoctoral Traineeship Award, Andrew R. Jamieson, has completed 17 out of the 20 required courses towards the Ph.D. degree in Medical Physics. The remaining three courses will be finished by 2010 Fall, at the latest. Some recent courses completed within the past year include Machine Learning, Anatomy of the Body, Health Physics, and a teaching assistantship in Computer Vision and Image Processing.

Research Accomplishments

1. Designed and Successfully Executed Proof of a Principle Grid-based Breast CADx Images Analysis Work-flow using Swift Script.

We designed and executed a pilot study to utilize large scale parallel grid computing to harness the nationwide cluster infrastructure for optimization of medical image analysis parameters. A previously developed CAD scheme for mass lesions in mammography was ported onto the grid computing environment by wrapping the algorithm code with the Swift script workflow language. The CAD scheme was then configured into a parallelizable workflow by the grid-software. The workflows were executed using two test clusters (in Santa Monica, CA and Chicago, IL) consisting of over 220 dual-CPU nodes combined. Using the grid-environment workflow, parameter sweeps were conducted for lesion segmentation settings based on radial-gradient-index (RGI) methods. Specifically, the Gaussian width (GW) used in initially filtering lesion images for segmentation was varied by increments of 1 mm from 1 to 60 mm. For each GW sweep the entire 850 biopsy-proven mass lesion database (411 benign, 439 malignant) was analyzed. In each, 29 different mathematical descriptor features were calculated, followed by feature selection and merging with linear discriminate analysis. Diagnostic performance was estimated by ROC analysis by calculating AUC (from PROPROC) values based on both individual features alone, and merged. For merged classifiers, AUC values were found using round-robin case-by-case removal and replacement. Among the resulting , computation jobs requiring over 30 CPU hours on a single lab computer were completed in approximately 35 minutes in this preliminary study. Merged AUC values increased from 0.50 (std.err.=0.018) at GW of 1mm with, to 0.81 (std.err.=0.015) at 10mm GW, with relative plateaus across the rest of the parameter space to 60mm.

In general, the parameter space sweep in GW identified trends in individual feature performance as well as merged results. Large scale, computationally intensive image analysis can be carried out in a timely fashion, feasible for expedited experimental discovery, as well as for more thorough future statistical analysis.

See **Appendix A** for Poster summarizing the work-flow and data found.

2. Investigation of Dimension Reduction(DR) in Place of Feature Selection Breast CADx

Recently developed unsupervised non-linear dimension reduction (DR) and data representation techniques were applied to computer-extracted breast lesion feature spaces across three separate imaging modalities: ultrasound (US) with 1126 cases, dynamic contrast enhanced-magnetic resonance imaging (DCE-MRI) with 356 cases, and full-field digital mammography (FFDM) with 245 cases. Two methods for non-linear DR were explored: Laplacian Eigenmaps of Belkin and Niyogi,[1] and t-distributed stochastic neighbor embedding (t-SNE) of van der Maaten and Hinton.[2] These methods attempt to map originally high-dimensional feature spaces to more human interpretable lower-dimensional spaces while preserving both local and global information.

Due in part to the ever-growing demand of data driven science, in recent years much interest has emerged in developing techniques for discovering efficient representations of large-scale complex data.[5] Conceptually the goal is to discover the intrinsic structure of the data and adequately express this information in a lower dimensional representation. Classically, the problem of DR and data representation has been approached by applying linear transformations such as the well-known Principal Component Analysis (PCA) or more general Singular Value Decomposition (SVD). [6,7] Interestingly, despite PCA's age, only recently has this method been considered for the specific application to CADx feature space reduction.[8] In this particular breast ultrasound study, while no significant boosts in lesion classification performance were discovered, PCA was found to be a suitable substitute in place of more computationally intensive and cumbersome feature selection methods.[8] This efficient lower dimensional PCA data representation, i.e. linear combinations of the original features accounting for the maximum global variance decomposition in the data, proved capable of capturing sufficient information for robust classification. However, PCA is not capable of representing higher order, non-linear, local structure in the data.

The goal of recently proposed non-linear data reduction and representation methods focuses on this very problem. [1,2] The present methods of interest to this study, Laplacian Eigenmaps and t-Distributed Stochastic Neighbor Embedding (t-SNE), offer two distinct approaches for explicitly addressing the challenge of capturing and efficiently representing the properties of the low dimensional manifold on which the original high-dimensional data may lie. Previous studies have investigated other non-linear DR techniques, including self-organizing maps (SOMs) and graph embedding, for breast cancer in the context of biomedical image signal processing[9,10], as well as for a breast cancer BIRADs database clustering[11]. To our knowledge the relationship between breast CADx performance and these non-linear feature space DR and representation have yet to be properly investigated. These new techniques may contribute two key enhancements to current CADx schemes.

- 1) A principled alternative to feature selection. Both methods explicitly attempt to preserve as much structure in the original feature space as possible, and thus require no need to assumingly force exclusion of features from the original set, and hence unnecessary loss of image information.
- 2) A more natural and sparse data representation that immediately lends itself to generating human-interpretable visualizations of the inherent structures present in the high-dimensional feature data.

3. Evaluation of the performance Dimension Reduction(DR) in Place of Feature Selection Breast CADx

For the high-dimensional feature spaces ,DR methods were tested across all modalities for a range of lower target dimensions and user-defined algorithm parameters. We evaluated the classifier performance using the area under the Receiver Operating Curve ROC curve (AUC) via the non-

parametric Wilcoxon-Mann-Whitney statistic, as calculated using the PROPROC software. Statistical uncertainty in classification performance due to finite sample sizes was estimated by implementing 0.632+ bootstrapping methods for training and testing the classifiers. Additionally, we computed the 95% empirical bootstrap confidence intervals on AUC values as estimated by no less than 500 bootstrap case set re-samplings. In all values reported, the sampling was conducting on a *by lesion* basis, as there may be multiple images associated with each unique lesion. In this regard, during classifier testing, the set of classifier outputs associated with a unique lesion were averaged to produce a single value. For the supervised feature selection methods (Automatic-relevance determination (ARD) and linear step-wise), feature selection was conducted, up to the specified number of features, on each bootstrapped sample set. Notably, the more general Markov Chain Monte Carlo based Bayesian artificial neural network (MCMC-BANN) was coupled with both the non-linear ARD and linear-based feature selection methods, while the linear discriminant analysis (LDA) was only with the LSW selection. As some of the calculations are computationally intensive, particularly the t-SNE mappings and MCMC-BANN training for the larger US data set, a 256-CPU shared.

Results are summarized in depth in **Appendix B**.

4. Investigation Breast CADx Feature Data Representation and Visualization with Non-Linear Local Geometry Preserving Dimension Reduction Methods

Please see **Appendix B** and associated figures for sample breast image CADx feature data visualizations.

5. Investigation of Manifold Regularization for Breast CADx using Unlabeled Image Data

Supervised classification, which embodies the traditional role of CADx, critically requires that the “truth” or true biological disease status, for instance “malignant” or “benign”, must be known for each case image during algorithmic training. However, accomplishing this data assembly step is often the most resource expensive component of conducting CADx research, and usually acts as a severely limiting factor. While efforts will continue to streamline the gathering of pathological and radiological information associated with each clinical case, in most research contexts, a relative abundance of readily available unlabeled data may persist. From a practical standpoint it is wasteful to completely discard this information. Although the “truth” may be unknown, these unlabeled cases still contain potentially useful image information. In particular, the unlabeled image data can be regarded as an additional sample drawn from the underlying marginal probability distribution characteristic of the combined class-categories, i.e. both “malignant” and “benign”. A large enough unlabeled data sample may provide sufficient knowledge of the inherent structure of the underlying marginal image distribution to guide the improved design of supervised classification on labeled cases. More simply stated, the unlabeled data may help “regularize” the training of CADx algorithms, and consequently improve clinical performance.

Recently developed unsupervised non-linear dimension reduction and data representation techniques, specifically Laplacian Eigenmaps (Belkin and Niyogi) and t-SNE (van der Maaten and Hinton), offer a principled approach for integrating unlabeled and labeled image feature data. Thus, the purpose of this study is to investigate the potential these methods have for leveraging unlabeled data information towards the design of more robust/stable CAD classification algorithms. This

preliminary study focuses within the context of diagnosing breast mass lesions for ultrasound.

The canonical methodology for training/testing CADx algorithms usually consists of the following distinct steps: 1) segmentation and feature extraction, 2) feature selection, 3) merged feature- supervised classifier training, and 4) performance evaluation. The incorporation of unlabeled data into the CADx algorithm training for regularization can be accomplished by employing unsupervised *dimension reduction* in place of explicit feature selection in step two. For example, in the ultrasound dataset used in this study, up to 81 features are extracted from the lesion images. Features may be extracted for both labeled and unlabeled cases (truth known and unknown). Instead of using supervised feature selection (such as automatic relevance determination), which is dependent exclusively upon the labeled cases, unsupervised dimension reduction can be used to map the high dimensional feature vectors, including the unlabeled feature data, into a lower dimensional representation. The reduced dimension mapped output for the labeled cases may then be used as input into supervised classification training. Crucially, the unlabeled cases have exerted influence on the relative mapping of the labeled cases used to train the classifier. Ideally, this influence serves as a regularizing force, leading to more robust performance on novel cases. Additionally, the reduced dimension representations may be amenable to useful visualizations in 2D and 3D.

Appendix C contains graphic illustrations of the concept of dimensionality reduction as well as the incorporation of unlabeled into the CADx training algorithm.

KEY RESEARCH ACCOMPLISHMENTS

- Designed and executed Swift script enabled Grid computing work-flows, and managed to display clear proof of principle for large scale, parallel breast image analysis.
- Made effective use of the new 256 CPU SIRAF Shared Computing Facility at the University of Chicago Dept. of Radiology both as a test-bed for large scale parallel job submission and workflow management and to rapidly conduct new Multi-Modality Breast Image CADx research, culminating in a peer-reviewed journal article submission. Estimated Total CPU time used: ~100,000 to 300,000 hours.
- Investigated the use of cutting edge data-analysis/mining techniques as applied to Ultrasound, FFDM, and DCE-MRI Breast Image Feature Space Analysis for CADx , specifically, dimension reduction and data representation techniques (t-SNE and Laplacian Eigenmaps) for high dimensional data spaces. These methods allow for an alternative to traditional feature selection methods. Using the high-throughput cluster computing capabilities, performance metrics and intensive statistical cross-validation (0.632+ bootstrap and ROC analysis for AUC performance) were performed to gain understanding of the new techniques potential versus previous Breast CADx methodologies. Results indicate the ability to rival or exceed previous CADx performance.
- The dimensional reduction and data representation techniques also were shown to provide rich visualization output for human interpretation of the complex breast image feature space geometry.
- Additionally, the promising findings and have motivated a number of new research avenues. Most significantly, the incorporation and principled use of "unlabeled" (truth-unknown/non-biopsy proven) image data for the training of CADx algorithms. Specifically, the unsupervised dimension reduction techniques can use the feature space geometric structure to help regularize algorithmic training.

REPORTABLE OUTCOMES

Conference Presentations and Abstracts

- **A.R. Jamieson**, M L Giger, M Wilde, L Pesce, I Foster, “Grid-Computing for Optimization of CAD,” *Poster*, 50th Assembly and Annual Meeting of American Association of Physicist in Medicine (AAPM), Houston, Illinois, USA, July 2008
- **A.R. Jamieson**, ML Giger, L. Pesce “Regularized Training of CADx Algorithms with Unlabeled Data Using Dimension Reduction Techniques,” *Accepted talk*. 95nd Assembly and Annual Meeting of Radiological Society of North America, Chicago, Illinois, USA, December 2009.
- **A.R. Jamieson**, M L Giger, et. al. “ Exploring Non-Linear Feature Space Dimension Reduction and Data Representation in Breast CADx”, *Accepted talk*. 51st Assembly and Annual Meeting of American Association of Physicist in Medicine (AAPM) Anaheim CA, USA, July 2009

Peer-reviewed Journal Papers

A.R. Jamieson, M. L. Giger, et. al. “Exploring Non-Linear Feature Space Dimension Reduction and Data Representation in Breast CADx with Laplacian Eigenmaps and t-SNE”, *Med. Phys.*, (*Accepted, in revision*), 2009.

CONCLUSIONS

Overall, the past year proved to be highly productive for the recipient of the Predoctoral Traineeship Award, in terms of both research output achieved, leading to the continued personal development as an independent investigator, as well as the intensive didactic training and education. The recipient has completed all but three classes (out of the 20 required) and will finish the required course load by Fall 2010, at the latest. The courses completed, both core Medical Physics and research oriented theoretical based electives, have helped to guide important research decisions.

The initial research effort completed during this first year has not only produced encouraging results but also laid an excellent foundation for beginning large scale deployment of the experimental techniques onto the Grid environment. In addition to successfully deploying a proof of principle Grid run, we have used the local 256-CPU cluster effectively to gain research direction. Specifically, we have managed to identify and learn to use newly developed data analysis methods which can powerfully enhance the proposed objective of uncovering important information from large databases of breast cancer image data. These new methods will allow us to interpret the nature of the underlying structure associated with image data. By investigating the structure of the multi-modality breast image data, we can then correlate the findings with other biological and genomic data towards maximizing the overall impact of such systems for future clinical deployment.

Furthermore, as mentioned in the report above, through investigation of the new dimensional reduction and data representation techniques, new capabilities have been identified for the use of "unlabeled" image feature data. This is particularly of interest, as the vast majority of clinically accumulated data is never analyzed for proof of biologic origin and pathology ("labeled"). Thus, introduction of these new methods to our research is exciting, as not only will more data ("unlabeled") be available to incorporate into our algorithmic development, but also, ample opportunity to make full use of the large scale computing power of the Grid.

REFERENCES

- [1] M. Belkin, and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," *Neural Comput.* **15**, 1373--1396 (2002).
- [2] L. van der Maaten, and G. Hinton, "Visualizing Data Using T-SNE," *J. Mach. Learn. Res.* **9**, 2605, 2579 (2008).
- [3] B. Efron, and R. Tibshirani, "Improvements on Cross-Validation: The .632+ Bootstrap Method," *J. Am. Stat. Assoc.* **92**, 548-560 (1997).
- [4] Tabar L, Yen MF, Vitak B, Tony Chen HH, Smith RA, and Duffy SW: Mammography service screening and mortality in breast cancer patients: 20-year follow-up before and after introduction of screening. *Lancet* 361: 1405-10, 2003.
- [5] Y. Wang, D.J. Miller, and R. Clarke, "Approaches to Working in High-Dimensional Data Spaces: Gene Expression Microarrays," *Br. J. Cancer* **98**, 1023-1028 (2008).
- [6] H. Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components," *J. Educ. Psychol.* **24**, 498-520 (1933).
- [7] M. Kirby, *Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns* (John Wiley & Sons, Inc., New York, 2000).
- [8] K. Drukker, N.P. Gruszauskas, and M.L. Giger, "Principal component analysis, classifier complexity, and robustness of sonographic breast lesion classification", *Medical Imaging 2009: Computer-Aided Diagnosis*, edited by M.Giger and N. Karssemeijer (2009), vol. 7260, *Proc. in SPIE*, pp. 72602B-6.
- [9] C. Varini, A. Degenhard, and T.W. Nattkemper, "Visual Exploratory Analysis of DCE-MRI Data in Breast Cancer by Dimensional Data Reduction: A Comparative Study," *Biomed. Signal Process. Control* **1**, 56-63 (2006).
- [10] A. Madabhushi, P. Yang, M. Rosen, and S. Weinstein, "Distinguishing Lesions from Posterior Acoustic Shadowing in Breast Ultrasound Via Non-Linear Dimensionality Reduction," *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **1**, 3070-3073 (2006).
- [11] M.K. Markey, J.Y. Lo, G.D. Tourassi, and C.E. Floyd, "Self-Organizing Map for Cluster Analysis of a Breast Cancer Database," *Artif. Intell. Med.* **27**, 113-127 (2003).

APPENDICES

- Appendix A: *POSTER*: **A.R. Jamieson**, M L Giger, M Wilde, L Pesce, I Foster, “Grid-Computing for Optimization of CAD,” *Poster*, 50th Assembly and Annual Meeting of American Association of Physicist in Medicine, Houston, Illinois, USA, July 2008
- Appendix B: *PAPER*: **A.R. Jamieson**, M. L. Giger, et. al. “Exploring Non-Linear Feature Space Dimension Reduction and Data Representation in Breast CADx with Laplacian Eigenmaps and t-SNE”, *Med. Phys.* (*Accepted, in revision*), 2009.
- Appendix C: *FIGURE*: **A.R. Jamieson**, ML Giger, L. Pesce “Regularized Training of CADx Algorithms with Unlabeled Data Using Dimension Reduction Techniques.” *Accepted talk* 95nd Assembly and Annual Meeting of Radiological Society of North America, Chicago, Illinois, USA, December 2009.

Purpose

Breast Cancer & CAD

Grid Computing

Materials and Methods

mathematical descriptor features were calculated.

A.R. Jamieson¹, M.L. Giger¹, L. Pesce¹, M. Wilde^{2,3}, I. Foster^{2,3}

³ The Computation Institute, University of Chicago

RGI-Segmentation

$$h(x, y) = f(x, y)N(x, y; \mu_x, \mu_y, \sigma_c^2)$$

$$M_i = \{(x, y) : h(x, y)$$

$$RCI = \frac{1}{\sum_{(x,y) \in M_k} \|\vec{g}(x,y)\| \cdot \vec{r}(x,y)}$$

CAD-Grid Work Flow



This study was supported in part by NIH grants T32 EB002103 and P50 CA125183. Additionally, Open Science Grid and Teragrid provided computing resources. Conflict of interest statement: M. L. Giger is a stockholder and receives royalties from R2Hologic.

Using the Grid computing environment, the Swift workflow executed start to finish in under an hour for the

restrictively small GW constraints (below 12mm) used for RGI mass lesion segmentation.



The most relevant result of this preliminary effort is the

analysis can be performed in a timely fashion, feasible for expedited experimental discovery and validation.

[1] Huo Z, et. al. Academic Radiology 5: 155-168, 1998

- [5] Pesce L, Metz C. Academic Radiology, 14(7): 814-829, 2007

Appendix B

MANUSCRIPT SUBMITTED TO *MEDICAL PHYSICS* JOURNAL

Title:

Exploring Non-Linear Feature Space Dimension Reduction and Data Representation in Breast CADx with Laplacian Eigenmaps and t-SNE

Authors: Andrew R. Jamieson, Maryellen L. Giger, Karen Drukker, Hui Li, Yading Yuan, and Neha Bhooshan

Department of Radiology, University of Chicago, Chicago, Illinois 60637

Abstract:

In this preliminary study, recently developed unsupervised non-linear dimension reduction (DR) and data representation techniques were applied to computer-extracted breast lesion feature spaces across three separate imaging modalities: ultrasound (US) with 1126 cases, dynamic contrast enhanced-magnetic resonance imaging (DCE-MRI) with 356 cases, and full-field digital mammography (FFDM) with 245 cases. Two methods for non-linear DR were explored: Laplacian Eigenmaps of Belkin and Niyogi,¹ and t-distributed stochastic neighbor embedding (t-SNE) of van der Maaten and Hinton.² These methods attempt to map originally high-dimensional feature spaces to more human interpretable lower-dimensional spaces while preserving both local and global information. The properties of these methods as applied to breast computer-aided diagnosis (CADx)

were evaluated in the context of malignancy classification performance as well as in the visual inspection of the sparseness within the two- and three-dimensional mappings. Classification performance was estimated by using the reduced dimension mapped feature output as input into both linear and non-linear classifiers: Markov Chain Monte Carlo based Bayesian artificial neural network (MCMC-BANN) and linear discriminate analysis (LDA). The new techniques were compared to previously developed breast CADx methodologies, including Automatic Relevance Determination (ARD) and linear step-wise (LSW) feature selection, as well as a linear DR method based on Principal Component Analysis (PCA). Using ROC analysis and 0.632+ bootstrap validation, 95% empirical confidence intervals were computed for the each classifier's AUC performance. Results: In the large US dataset, sample high performance results include, $AUC_{0.632+} = 0.88$ with 95% empirical bootstrap interval [0.787;0.895] for 13 ARD selected features and $AUC_{0.632+} = 0.87$ with interval [0.817;0.906] for 4 LSW selected features compared to 4D t-SNE mapping (from the original 81D feature space) giving $AUC_{0.632+} = 0.90$ with interval [0.847;0.919], all using the MCMC-BANN. Conclusions: Preliminary results appear to indicate capability for the new methods to match or exceed classification performance of current advanced breast lesion CADx algorithms. While not appropriate as a complete replacement of feature selection in CADx problems, DR techniques offer a complementary approach which can aid elucidation of additional properties associated with the data. Specifically, the new techniques were shown to possess the added benefit of delivering sparse lower-dimensional representations for visual interpretation, revealing intricate data structure of the feature space.

Keywords: non-linear dimension reduction, computer-aided diagnosis, breast cancer, Laplacian Eigenmaps, t-SNE

I. Introduction

Radiologic image interpretation is a complex task. A radiologist's expertise, developed only with exhaustive training and experience, rests in their ability for extracting and meaningfully synthesizing relevant information from a medical image. However, even under idealized image acquisition conditions, precise conclusions may not be possible for certain radiologic tasks. Thus, computer aided diagnosis (CADx) systems have been introduced in a number of contexts in an attempt to assist human interpretation of medical images.³ A relatively well-developed clinical application for which computerized efforts in radiological image analysis have been studied is the use of CAD in the task of detecting and diagnosing breast cancer.⁴⁻¹⁰ Similar to the radiologist's task, a computer algorithm is designed to make use of the highly complicated breast image input data, attempting to intelligently reduce image information into more interpretable and ultimately clinically-actionable output structures, such as an estimate of the probability of malignancy. Understanding how to optimally make use of the enormity of the initial image information input and best arrive at the succinct conceptual notion of "diagnosis" is a formidable challenge. Although there may be any number of various operations/transformations involved in arriving at this high-level end output, whether in the human brain or *in silico*, two common critical pursuits are proper data representation and reduction. The current study aims to explore the potential enhancements offered to breast mass lesion CADx algorithms through the application of two recently-developed dimensionality reduction and data representation techniques, Laplacian Eigenmaps and t-distributed stochastic neighbor embedding (t-SNE).^{1,2}

II. Background

II.A. Current CADx Feature Representation

Restricted by limited sample datasets, computational power, and lack of complete theoretical formalism, image-based pattern recognition and classification techniques often tackle the objective task at hand by substantially simplifying the problem. Traditionally, breast CADx systems employ a two pronged approach, first, image pre-processing and feature extraction, and second, classification in the feature space, either by unsupervised methods, supervised methods, or both. A review of past and present CADx methods employed can be found in referenced articles referenced.^{3,11} Often, instead of attempting to make use of the complete image¹², CADx typically condenses image information down to a vector of numerical values, each representative of some attribute of the image or lesion present in the image. One can consider this first data reduction step as “perceptual” processing, meaning that at this stage the algorithm’s goal is to isolate and “perceive” only the most relevant components of the original image that will contribute towards distinguishing between the target classes (e.g., malignant or benign). One of the steps in eliminating unnecessary image information is lesion margin segmentation.^{5,13} Typically, features, such as those extracted from the segmented lesion, are heuristic in nature and mimic important human identified aspects of the lesion. However more mathematical and abstract feature quantities may also be calculated that may represent information visually imperceptible to the unaided eye. While the use of data from a segmented lesion introduces bias into the algorithm’s task as a whole, this “informed” bias allows for the efficient removal of much unnecessary image data, for instance normal background breast tissue. From here the second main component of the CADx algorithm falls usually into the context of the well-formalized canonical problem found in statistical pattern recognition for classification^{14,15}.

After the first CADx phase of feature extraction, each high-dimensional image in the sample set is now reduced to a single vector in a lower-dimensional feature space. However, due to the finite size of image sample data, if too many features are examined simultaneously, regions

containing a low density of points in the feature space will exist, resulting in statistically inconclusive classification ability. This dilemma is affectionately termed the “curse of dimensionality.”¹⁶ Thus, a further reduction in the full feature space is required for a practically useful data representation. This aspect is a major concern of the second component of traditional CADx schemes, and is succinctly known as “feature selection”. Much literature has been generated on this subject matter in the explicit context of improving CADx performance¹⁷⁻¹⁹ Some CADx schemes may employ only 4-5 features maximum, in which case, feature selection may not be necessary, since the dataset sample size, even for relatively smaller sizes, may be sufficiently large to avoid over-training classifiers. However, it is reasonable to imagine CADx researchers interested in testing hundreds of potential features. In either case, when appropriately coupled with a well-regularized supervised classification method, the ultimate objective of features selection is to discover the “optimal” data representation, or sub-set of features for robustly maximizing the desired diagnostic task performance. That is, the method attempts both to mimic and to maximize the theoretical upper bound or ideal observer performance possible over the sampled joint probability distribution of the selected features. While this step is critical, finding such a sub-set is non-trivial and may also be highly dependent on the specific characteristics of the sample data. Developed techniques in feature selection for CADx range from simpler linear methods, such as those based on linear discriminate analysis (LDA), to non-linear and more sophisticated Bayesian-based, such as the use of Bayesian Artificial Neural Networks (BANN) and Automatic Relevance Determination (ARD), to random-search stochastic methods such as genetic algorithms as well as information theoretic techniques^{17,19-21}.

The most striking quality of the methods mentioned above, in the context of CADx, is that during feature selection, some features are completely removed from the final classification scheme, and hence image information is either explicitly or implicitly discarded altogether. However, while

removing out all the information associated with a specific feature not selected, by selecting a smaller sub-set of individual features, what is gained is greater immediate human interpretability. Specifically, the isolated groups of features may have clear physical or radiological meanings and thus may be of interest to investigators or radiologists for understanding how these characteristics relate to the ability to distinguish class categories (malignant, benign, cyst, etc.). To this end, in order to interpret the nature of the feature space and attempt to identify characteristic trends, one may visually inspect plots displaying single features or attempt to capture synergistic qualities between two or three features simultaneously. Above three dimensions, as it becomes non-trivial to interpret the structure of the feature space, often instead, the use of a metrics such as the ROC curve and/or AUC based on output from the decision variable of a trained merged feature classifier are used to interrogate the quality of the higher dimensional feature spaces.

As such, beyond identifying which feature or features appear to hold classification utility, current CADx methods offer little theoretical/formal guidance in a recovering understanding of the inherent data structure represented by the higher dimensional feature spaces.

II.B. Proposed Feature Space Representation and Reduction for CADx

Due in part to the ever-growing demand of data driven science, in recent years much interest has emerged in developing techniques for discovering efficient representations of large-scale complex data.²² Conceptually the goal is to discover the intrinsic structure of the data and adequately express this information in a lower dimensional representation. Classically, the problem of dimension reduction(DR) and data representation has been approached by applying linear transformations such as the well-known Principal Component Analysis (PCA) or more general Singular Value Decomposition (SVD).^{23,24} Interestingly, despite PCA's age, only recently has this method been considered for the specific application to CADx feature space reduction.²⁵ In this

particular breast ultrasound study, while no significant boosts in lesion classification performance were discovered, PCA was found to be a suitable substitute in place of more computationally intensive and cumbersome feature selection methods.²⁵ This efficient lower dimensional PCA data representation, i.e. linear combinations of the original features accounting for the maximum global variance decomposition in the data, proved capable of capturing sufficient information for robust classification. However, PCA is not capable of representing higher order, non-linear, local structure in the data.

The goal of recently proposed non-linear data reduction and representation methods focuses on this very problem.^{1,2} The present methods of interest to this study, Laplacian Eigenmaps and t-Distributed Stochastic Neighbor Embedding (t-SNE), offer two distinct approaches for explicitly addressing the challenge of capturing and efficiently representing the properties of the low dimensional manifold on which the original high-dimensional data may lie. Previous studies have investigated other non-linear DR techniques, including self-organizing maps (SOMs) and graph embedding, for breast cancer in the context of biomedical image signal processing^{26,27}, as well as for a breast cancer BIRADs database clustering.²⁸ To our knowledge the relationship between breast CADx performance and these non-linear feature space DR and representation have yet to be properly investigated. These new techniques may contribute two key enhancements to current CADx schemes.

1. A principled alternative to feature selection. Both methods explicitly attempt to preserve as much structure in the original feature space as possible, and thus require no need to assumingly force exclusion of features from the original set, and hence unnecessary loss of image information.
2. A more natural and sparse data representation that immediately lends itself to generating human-interpretable visualizations of the inherent structures present in the high-

dimensional feature data.

It is important to note that by employing DR on CADx feature spaces, one surrenders, to a varying extent, the ability to immediately interpret the physical meaning of the embedded representation. Yet, critically, this is a necessary and fundamental trade-off, as the conceptual focus is shifted to a more holistic approach, specifically, that of discovering an efficient lower dimensional representation of the intrinsic data structure. The core tenant of such an unsupervised approach is to limit assumptions imposed on the data. This major shift in philosophy regarding the original high dimensional feature space embodies the notion, “let the data speak for itself.” It seems reasonable to assume that if supervised classifiers are capable of uncovering sufficient data structure in the extracted feature space for producing adequate classification performance, then such principled local geometry preserving reduction mappings should reveal structural evidence corroborating such findings.

II.C. Outline of Evaluation for Proposed Methods

The primary objective of this study is to evaluate the classification performance characteristics of breast lesion CADx schemes employing the Laplacian Eigenmap or t-SNE DR techniques in place of previously developed feature-selection methods. Secondly, and more qualitatively, we aim to investigate and gain insight into the properties of sample visualizations representative of lower-dimensional feature space mappings of high-dimensional breast lesion feature data. Additionally, the feasibility and robustness of these non-linear reduction methods for CADx feature space reduction are tested across three separate imaging modalities: ultrasound (US), dynamic contrast enhanced MRI (DCE-MRI), and full-field digital mammography (FFDM), having case sets of 1126 case, 356 cases, and 245 cases, respectively.

III. Methods

III.A. Dataset

All data characterized in this study consists of clinical breast lesions presented in images acquired at the University of Chicago Medical Center. Lesions are labeled according to the truth known by biopsy or radiologic report and collected under HIPAA-compliant IRB protocols. Furthermore, the breast lesion feature datasets were generated from previously developed CADx algorithms at the University of Chicago. For a review of these techniques see Giger, Huo, Kupinski for X-ray mammography, Drukker, for US, and Chen for DCE-MRI.^{4-11,29}

In each of the modalities, the lesion center is identified manually for the CADx algorithm, which then performs automated-seeded segmentation of the lesion margin followed by computerized feature extraction. Table 1 below summarizes the content of the respective imaging modality databases used, including the total number of initial lesion features extracted. Note that the mammographic imaging modality (FFDM) contains only two lesion class categories, malignant and benign. For ultrasound and DCE-MRI a more detailed sub-categorization is provided, including invasive carcinoma (IDC), ductal carcinoma *in situ* (DCIS), benign solid masses, and benign cystic masses. For clarity, this initial study only considers binary classification performance in the task of distinguishing between the more broad identity of malignant and benign (cancerous vs. non-cancerous). However, during qualitative inspection of the dimension reduced mappings, it will be of interest to re-introduce these distinctions for visualization purposes.

Modality	Total Number of Images	Number of Malignant Lesions	Number of Benign Lesions	Total Number of Lesion Features Calculated
----------	------------------------------	-----------------------------------	-----------------------------	--

US	2956	158	968 (401 mass / 567 cystic)	81
DCE-MRI	356	223 (151 IDC / 72 DCIS)	133	31
FFDM	735	132	113	40

Table 1. Feature Database Characteristics.

Geometric, texture, and morphological features, such as margin sharpness, were extracted across all modalities. Also, the DCE-MRI dataset includes kinetic features, and the US features include those related to posterior acoustic behavior.^{8,10} All raw extracted feature value datasets were normalized to zero mean and divided by the unit sample standard deviation. Due to page limitations, the details of each feature can be found in the referenced papers.^{4-11,29}

III.B. Classifiers

In our evaluation of the new DR techniques, we chose two types of classifiers: a relatively simple linear discriminant analysis (LDA) classifier and a more sophisticated non-linear, Bayesian artificial neural network, classifier (BANN).¹⁵ LDA is a well-known and commonly used linear classification method which will not be reviewed here, for reference and examples in breast lesion CADx see references.^{4,30,31} The BANN, as the name suggests, follows the usual multi-layer-perception, neural network design, but additionally employs Bayesian theory as a means of classifier regularization^{15,32}. The BANN has been shown to model the optimal ideal observer for classification given sufficient sample sizes as input for training.³³ The critical technical hurdle in implementing BANNs lies in accurately estimating posterior weight distributions, as analytical calculation is intractable. As such, either approximation or sampling based methods must be deployed in practice.³⁴ Markov Chain Monte Carlo (MCMC) sampling methods can be used to

directly sample from the full posterior probability distribution.³² We implemented a MCMC-BANN classifier using Nabney’s *Netlab* package for MATLAB.³⁵ The following network architecture, $k--(k+1)--1$, was used. That is, k input layer nodes (one for each of the k selected features), a hidden layer with $(k + 1)$ nodes, and a single output target as probability of malignancy. For each classifier trained, we generated at least 2000 MCMC samples of the weights’ posterior probability distribution. The mean value of the classification prediction (probability of malignancy) output from each of the different 2000 weight samples was used to produce a single classification estimate for new test input cases.

III.C. Explicit Supervised Feature Selection Methods

Two previously developed feature selection methods are considered in this paper for comparison, and include linear step-wise and ARD feature selection. These methods are used to identify a specific set of features for input into the classifier.

Linear Stepwise Feature Selection

Linear step-wise feature selection (LSW-FS) relies on linear discriminant-based functions. Beginning with only a single selected feature, multiple combinations of features are considered one at a time, by exhaustively adding, retaining, or removing each subsequent feature to the potential set of selected features. For each new combination, a metric, the Wilks’ lambda is calculated and a selection criterion based on F-statistics is used.¹⁷ The “F-to-enter” and “F-to-remove” used in this study were automatically adjusted to allow for the specified number of features desired for US, DCE-MRI, and FFDM feature selection. For examples of LSW-FS use in breast CADx references are provided.^{17,25,30}

Automatic Relevance Determination

A consequence of the BANNs is the possibility for joint feature selection and classification using Automatic Relevance Determination (ARD).^{15,32,34,35} ARD works by placing Bayesian hyper-priors, also known as hierarchical priors, over the initial prior distributions already imposed on the network weights connected to the input nodes. The “relevant” features are then discovered as estimates for the hyper parameters, which characterize the prior distributions over the respective input layer weights, are updated via Gibbs sampling giving the posterior hyper-parameter estimate. The magnitudes of the final, converged upon hyper-parameters are then used to indicate the relative utility of the respective feature input layer weights towards accomplishing the classification task. Thus, by way of the Bayesian regularization, ARD allows for one-shot feature selection and classifier design. Furthermore, a key advantage of ARD feature selection is its ability to identify important non-linear features coupled to the classification objective, due to the inherent non-linear nature of BANN.¹⁹ Due to these qualities, ARD-MCMC-BANN classifiers were also included for comparison in our study.

In this study we extend MCMC-BANN to incorporate ARD following the implementation of Nabney.³⁵ This methodology was previously investigated for breast feature selection and classification in DCE-MRI CADx.¹⁹ In our study, 1000 samples were calculated for the hyper-parameters beginning with a gamma hyper-prior distribution of mean parameter value equal to 3 and a shape parameter equal to 4.

III.D. Unsupervised Dimension-Reduction Feature Mappings

In comparison to the supervised feature-selection methods, three unsupervised DR methods were evaluated here; the latter two non-linear methods are offered as a novel application to the field of breast image CADx. The general problem of dimensionality reduction can be described

mathematically as: provided an initial set x_1, \dots, x_k of k points in R^l , discover a set y_1, \dots, y_k in R^m such that y_i sufficiently describes or “represents” the qualities of interest found in the original set x_i . In the context of breast lesion CADx feature extraction, the ideally lower dimensional mappings should aim to preserve and represent as much relevant structural information towards the task of malignancy estimation. It should be noted that DR still requires, in some sense, “feature selection,” meaning, one must specify the number of mapped dimensions to retain for the subsequent classification step. Ideally, methods designed to estimate intrinsic dimensionality of the data structure could be used to direct this choice.³⁶ However, proper evaluation of the integrity of such methods in this context is beyond the scope of this research effort. Thus, in approaching the problem from a more naïve perspective, as done here, focus is centered on gaining a general intuition for the overall major trends encountered.

Linear Feature Reduction: PCA

Mathematically, PCA is linear transformation which maps the original feature space onto new orthogonal coordinates. The new coordinates, or principal components (PC), represent ordered orthogonal data projections capturing the maximum variance possible, with the first PC corresponding to the highest global variance.^{23,24} Drukker, et al. used PCA as an alternative to feature selection for breast US CADx.²⁵

Non-linear Feature Dimension Reduction

As discussed in the introduction and background sections, the following two recently proposed DR and data representation methods are non-linear in nature and specifically designed to address the problem of local data structure preservation. Laplacian Eigenmaps and t-SNE offer highly distinct solutions to this problem.

i. Laplacian Eigenmaps

Drawing on familiar concepts found in spectral graph theory, Laplacian Eigenmaps, proposed by Belkin and Niyogi in 2002, use the notion of a graph Laplacian applied to a weighted neighborhood adjacency graph containing the original data set information.¹ This weighted neighborhood graph is regarded geometrically as a manifold characterizing the structure of the data. The eigenvalues and eigenvectors are computed for the graph Laplacian which are in turn utilized for embedding a lower dimensional mapping representative of the original manifold. Acting as an approximation to the Laplace Beltrami operator, the weighted graph Laplacian transformation can be shown, in a certain sense, to optimally preserve local neighborhood information.³⁷ Thus, the feature data considered in the reduced dimensional space mapping is essentially a discrete approximate representation of the natural geometry of the original continuous manifold.

As Belkin and Niyogi note, the algorithm is relatively simple and straightforward to implement. Additionally, the algorithm is not computationally intensive. For our largest dataset the mappings were computed within a few seconds using MATLAB code. Algorithm details as well as explanation of necessary input parameters for the implementation used here are provided below in section VIII.A of the Appendix.

It is important to note that there is no theoretical justification for how to choose the needed parameters for the algorithm. Thus, an array of parameter choices was evaluated in this study. Lastly, parts of the MATLAB code, related only to the implementation of the Laplacian Eigenmap, were modified from the publically available dimension reduction toolbox provided by Laurens van der Maaten of Maastricht University.³⁸

ii. t-Distributed Stochastic Neighbor Embedding (t-SNE)

The other non-linear mapping technique considered, t-Distributed Stochastic Neighbor Embedding (t-SNE) of van der Maaten and Hinton², approaches the dimension reduction and data reduction problem by employing entirely different mechanisms to the Laplacian Eigenmaps. t-SNE attacks DR from a stochastic and probabilistic-based framework. While requiring orders of magnitude more computational effort, such statistically-oriented approaches, provided they are well-conditioned, may potentially offer greater flexibility in certain contexts due in part by the lessening of potentially restrictive theoretical mathematical formalism. For these reasons the t-SNE method was considered as an interesting comparison alongside the Laplacian Eigenmap.

t-SNE is an improved variation on the original stochastic neighbor embedding (SNE) of Hinton and Roweis.³⁹ The basic idea behind SNE is to minimize the difference between specially defined conditional probability distributions that represent similarities, calculated for the data points in both the high and low dimensional representations. In particular, SNE begins by first computing the conditional probability $p_{j|i}$ given by

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|/2\sigma_i^2)} \quad \text{and} \quad q_{j|i} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)} \quad (1)$$

and $q_{j|i}$ in the lower dimensional space with $p_{i|i}$ and $q_{i|i}$ set to zero. These similarities express the probability that x_i (y_i) would select x_j (y_j) as its neighbor, resulting in high values for nearby points and lower values for distantly separated ones. The central assumption in SNE is that if the low-dimensional mapped points in Y space correctly model the similarity structure of its higher-dimensional counterparts in X , then the conditional probabilities will be equal. The summed Kullback-Leibler (KL) divergence is used to gauge how well $q_{j|i}$ models $p_{j|i}$. Using gradient descent methods, SNE minimizes a KL based cost function. Sampled points from an isotropic Gaussian with small variance centered at the origin are used to initialize the gradient decent. Updates are

made to the mapped space Y for each iteration. Additionally, the parameter σ_i of eq (1) must be selected. σ_i is the variance in the Gaussian centered on the high dimensional point x_i . Because of the difficulty in determining if an optimal σ_i exists, a user defined property called perplexity is used to facilitate its selection, defined by $Perp(P_i) = 2^{H(P_i)}$. Calculated in bits, $H(P_i)$ is the Shannon entropy over P_i

$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i} \quad (2)$$

During SNE, a binary search is performed to find the value of σ_i that produces a P_i with the user specified perplexity. Suggested typical settings range between 5 and 50.²

t-SNE introduces two critical improvements to SNE.² First, the gradient as well as cost function optimization is simplified by using symmetrized conditional probabilities to define the joint probabilities on P and Q (e.g. $p_{ij} = (p_{j|i} + p_{i|j})/2n$) and the minimizing cost over a single KL divergence as opposed to a sum,

$$C = \sum_i KL(P_i \parallel Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \Rightarrow C' = KL(P \parallel Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}. \quad (3)$$

Second, the distributional form of the low-dimensional joint probabilities is changed from a Gaussian, to the heavier tailed Student t-distribution with one degree of freedom. Roughly, this promotes a greater probability for moderately distanced data points in high dimensional space to be expressed by a larger distance in the low-dimensional map, thus more “faithfully” representing the original distance structure, and avoiding the “crowding problem.”² The new q_{ij} is defined as

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}} \quad (4)$$

After incorporating the altered q_{ij} , the final gradient for the cost function is given by

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1} \quad . \quad (5)$$

A step by step algorithm outline for t-SNE is provided in section VIII.B of the Appendix.

As recommended by Hinton and van der Maaten², PCA is first applied to the high-dimensional input data in order to expedite the computation of the pairwise distances. Lastly, as t-SNE was developed primarily for 2D and 3D data representation and visualization, it is important to note that the authors warn performance of t-SNE is not well understood for the general purpose of DR.² By applying t-SNE to the CADx feature reduction problem we hope to offer at least some empirical insight towards understanding its properties in such contexts. We used van der Maaten publicly available t-SNE MATLAB code and Intel processor optimized “fast_tsne” to generate the present data mappings⁴⁰.

III.E. Classifier Performance Estimation and Evaluation

The high-dimensional feature spaces DR methods were tested across all modalities for a range of lower target dimensions and user-defined algorithm parameters. We evaluated the classifier performance using the area under the Receiver Operating Curve ROC curve (AUC) via the non-parametric Wilcoxon-Mann-Whitney statistic, as calculated using the PROPROC software.⁴¹⁻⁴³ Statistical uncertainty in classification performance due to finite sample sizes was estimated by implementing 0.632+ bootstrapping methods for training and testing the classifiers.^{44,31} Additionally, we computed the 95% empirical bootstrap confidence intervals on AUC values as estimated by no less than 500 bootstrap case set re-samplings. In all values reported, the sampling was conducting on a *by lesion* basis, as there may be multiple images associated with each unique lesion. In this regard, during classifier testing, the set of classifier outputs associated with a unique lesion were

averaged to produce a single value. For the supervised feature selection methods (ARD and LSW), feature selection was conducted, up to the specified number of features, on each bootstrapped sample set. Notably, the more general MCMC-BANN was coupled with both the non-linear ARD and linear-based feature selection methods, while the linear LDA was only with the linear stepwise feature selection. As some of the calculations are computationally intensive, particularly the t-SNE mappings and MCMC-BANN training for the larger US data set, a 256-CPU shared computing resource cluster was employed to accomplish runs in a feasible time frame.

IV. Results

IV.A. Classification Performance.

MCMC-BANN and LDA classification performance is plotted as a function of the mapped or feature selected input space dimension for the three datasets, US, DCE-MRI, and FFDM, using the three different DR techniques, as well as the non-reduced selected features in Figure 1(a-f). Performance is characterized in terms of the 0.632+ bootstrapped AUC (left axis) and variability as gauged by the width of the empirical 95% bootstrap interval (right axis). The t-SNE perplexity was set to $Perp = 30$ and Laplacian Eigenmaps were generated with $Nearest Neighbor=45$ and $t=1.0$. Overall, the highest classification performance was attained by the largest sample-size US feature dataset with the DR-MCMC-BANN just slightly eclipsing the LDA, achieving approximately $AUC_{0.632+} \sim 0.90$, while the smaller DCE-MRI and FFDM feature data produced peaks around $AUC_{0.632+} \sim 0.80$. The variability in bootstrapped AUCs is also lowest for the large US dataset, hovering near ~ 0.07 as the number of inputs into the classifier is increased.

A few key observations can be made from the results regarding the use of DR. Primarily, the DR techniques, for both linear (PCA) and non-linear (t-SNE and Laplacian Eigenmaps), overall,

appear to at least match, or and in some cases exceed, explicit feature selection classification $AUC_{0.632+}$ performance. This is most evident when compared to the the ARD-FS coupled with the MCMC-BANN performance across all three imaging modalities, Figures 1(a), 1(c), 1(e) (left axis). Specifically, in all cases the DR methods exhibited a more rapid rise to peak $AUC_{0.632+}$ performance and remained higher than the ARD-based feature selection for all dimension input sizes. Additionally, compared to the ARD feature selection approach, the DR methods produced less variability in the bootstrap AUC. Figures 1(a),1(c),1(e) (right axis) substantially highlight this phenomenon. In particular, for the US data, the ARD-FS variability, being greater than all the DR methods, clearly trends downward as more features are selected for input; gradually approaching the DR variability levels, yet usually remaining higher. By comparison, save for a slight increase at 1D, the DR variability is relatively consistent from 2D to 13D.

However, when coupled with the LSW feature selection, the MCMC-BANN produced more competitive results against the DR performance. For example, for this MRI data set, except for 10D and 11D, the LSW-MCMC-BANN edged above all the DR based methods. Likewise, the use of the LSW feature selection with the MCMC-BANN resulted in substantially reduced variation in classifier performance compared to the ARD-FS. The LSW-MCMC-BANN variation nearly matched the DR output for both the US and MRI across all input dimensions. For the FFDM data, except for 2D-5D, the LSW-MCMC-BANN held close to the DR variation level.

The less complex, yet more stable LDA classifier, Figures 1(b),1(d),1(f)(left axis), produced different characteristic results. In all cases the LSW-feature selection performance was initially higher, however, as the dimension input space was increased, the DR methods became comparable. Expectedly, when coupled with the linear LDA, the highly-non-linear stochastic based t-SNE DR consistently underperformed. Turning to variation for the LDA, Figure 1(b), 1(d), 1(f) (right axis), the LSW-FS again exhibited different behavior from ARD-FS, in that, except for the smaller-case-

sized FFDM data, variability does not considerably fluctuate moving from 1D to 13D for both the LSW-FS and DR methods.

One manner by which to concisely analyze the performance characteristics of dimension-reduction/feature selection and classifiers designs for a particular dataset is to plot the bootstrap cross-validation AUC against the variability. An example is provided for the US feature dataset in Figure 2, with each point representing a different number of input dimensions. Data points located in the upper left corner indicate the most preferred performance qualities, i.e., higher classification performance and lower expected variability. Also provided in Figure 5, is a plot displaying classification results for both MCMC-BANN and LDA, in terms of the bootstrap AUC for the US data. Included within this plot are the empirical 95% confidence intervals to aid in gauging statistical significance for differences between estimated AUC values.

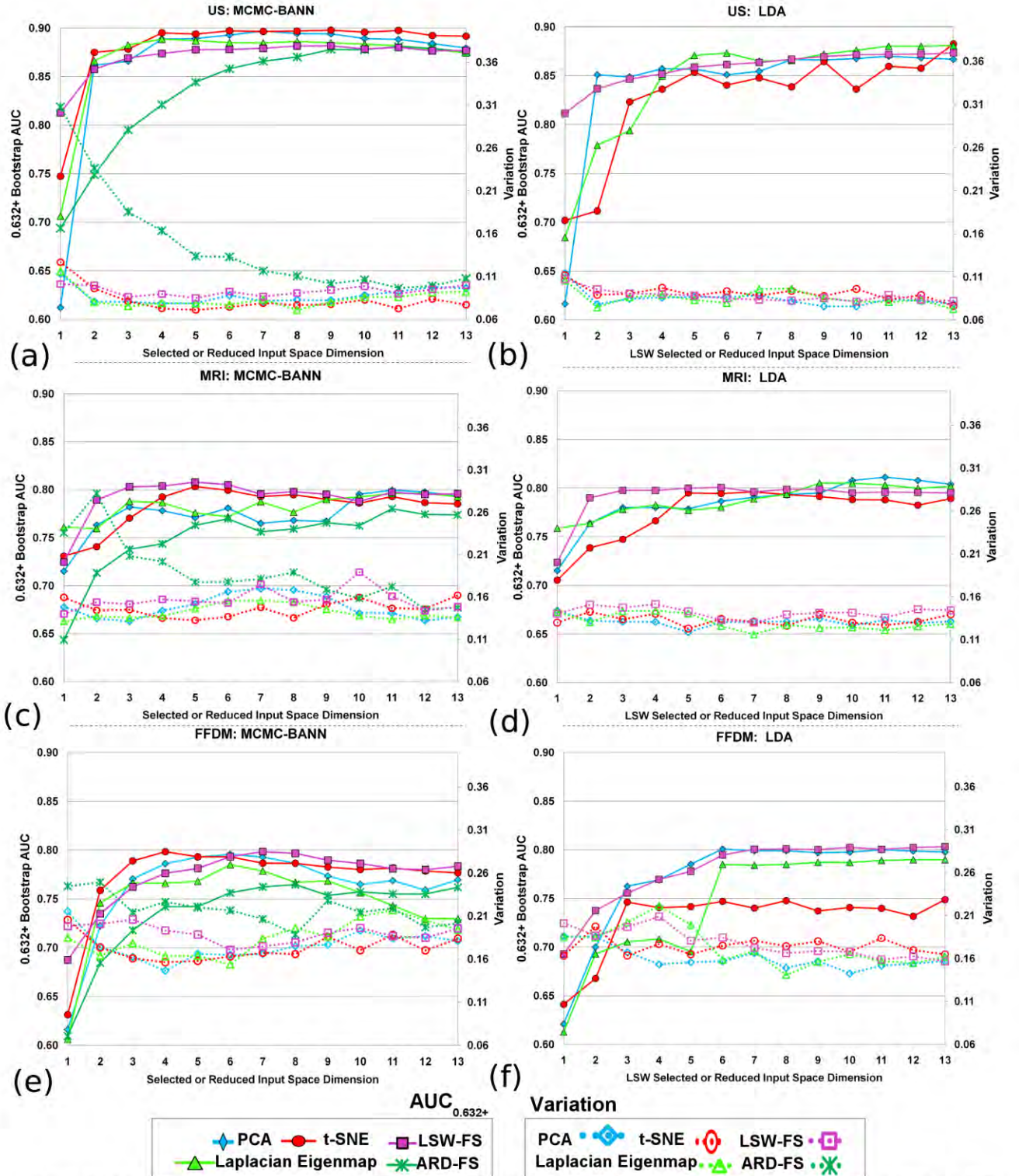


Figure 1. The $0.632+$ bootstrap area under the ROC curve (AUC) (left axis) and the variation as measured by the width of the 95% empirical bootstrap confidence intervals (right axis) versus the selected feature {ARD,LSW} or reduced representation {PCA,t-SNE,Laplacian Eigenmap} classifier input space dimension. (a) MCMC-BANN, (b) LDA, classifier performance on the originally 81 dimensional US feature dataset. (c) MCMC-BANN, (d) LDA classifier performance on the originally 31 dimensional DCE-MRI feature dataset. (e) MCMC-BANN, (f) LDA classifier performance on the originally 40 dimensional FFD feature dataset.

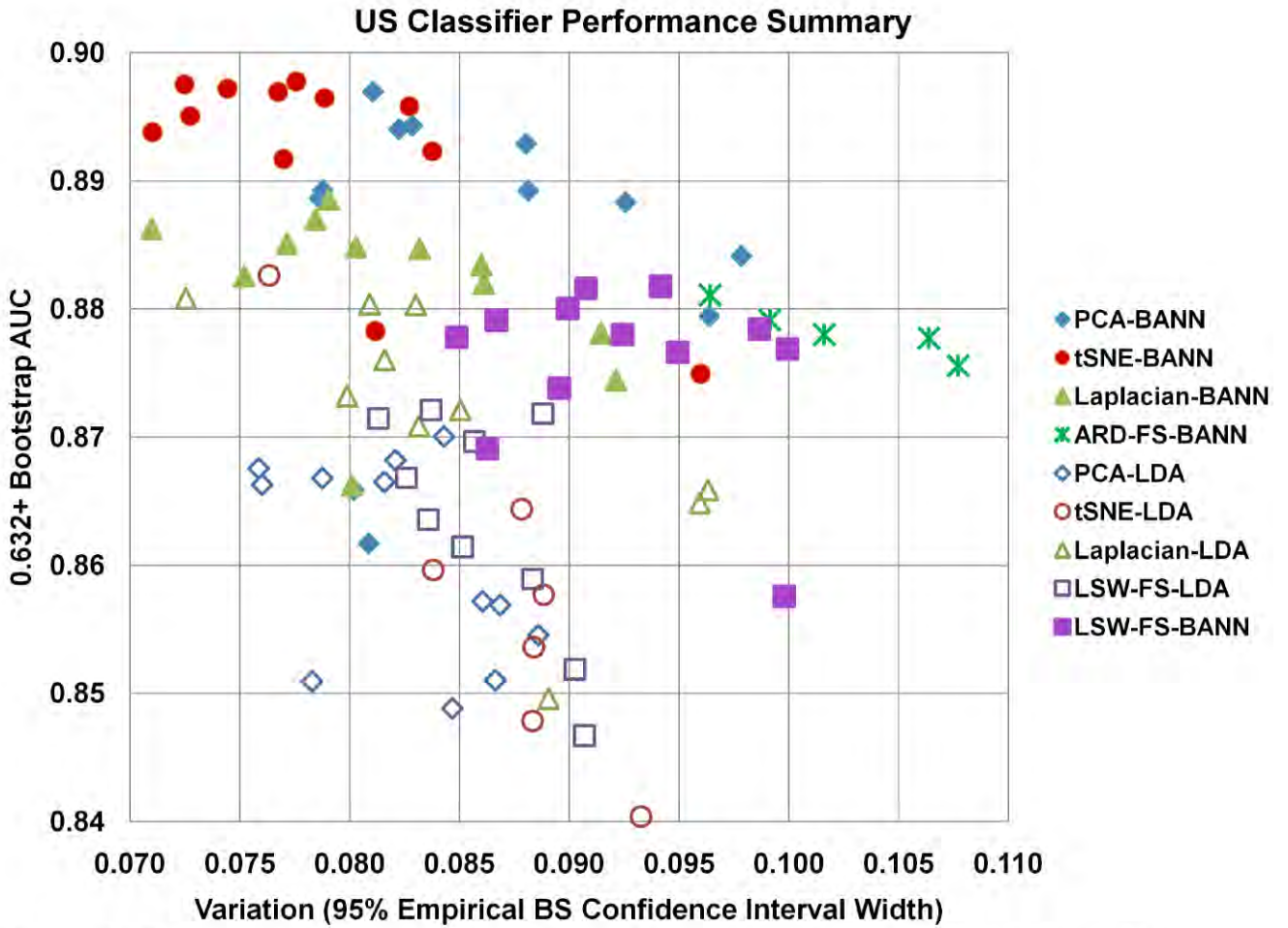


Figure 2. Summary of the classification performance on the 81 dimensional US feature dataset. The 0.632+ bootstrapped area under the ROC curve versus variability as gauged by the width of the 95% empirical bootstrap confidence intervals. Each point corresponds to a different input space dimension size. Points located in the upper left corner represent the highest expected AUC as well as least expected variation in performance due to sampling.

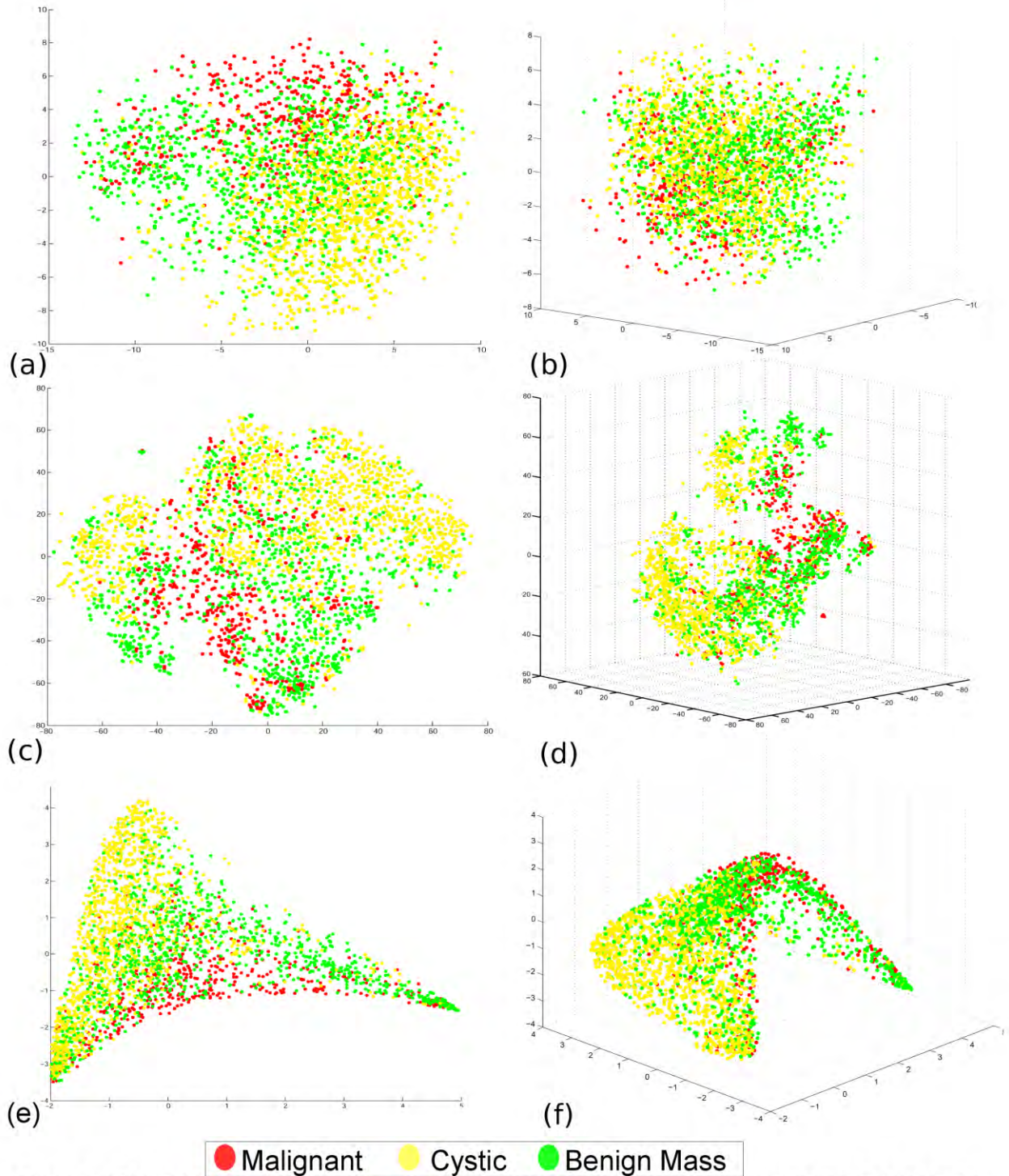


Figure 3. 2D and 3D visualizations of the unsupervised reduced dimension representations of the entire originally 81 dimensional breast lesion ultrasound feature dataset; green data points signifying benign lesions, red: malignant, and yellow: benign-cystic. (a) Visualization of linear reduction using PCA, first two principal components, (b) first three principal components, 3D PCA. (c) 2D and (d) 3D visualization of the non-linear reduction mapping using t-SNE. (e) 2D and (f) 3D visualization of the non-linear mapping using Laplacian Eigenmaps.

IV.B. 2D and 3D Visual Representations of Mappings

Due to the large sample size of the US feature data, a high density of points is produced (and hence the clearest delineation of structures) in the reduced dimension mapping representations. Figure 3(a-f) provides visual representations of the entire originally 81 dimensional US feature data mapped into 2D and 3D Euclidean space by the unsupervised PCA, t-SNE, and Laplacian Eigenmaps. The data points were subsequently colored to reflect the distribution of the lesions types (malignant tumor, benign lesion, cyst) with the reduced space.

Two key aspects are considered regarding the respective mappings: natural class separability and overall geometric traits characteristic of the represented structures, such as smoothness and sparsity. PCA is shown in Figures 3(a) and 3(b). Certain regions are potentially identifiable as being associated with a specific class (such as the dominance of cystic-benign points in the bottom right corner of the 2D plot), however, PCA generates a relatively homogeneous, nearly spherical distribution of points. Reflective of its mathematical basis, PCA representations provide primarily global information content, lacking the capability to represent rich local data structure. t-SNE generates a dramatically different type of low dimensional representation. As shown in figures 3(c) and 3(d), t-SNE produces a highly non-linear, jagged, and highly sparse data mapping. Many isolated “island-like” sub-groupings are identifiable in the t-SNE visual representations. As predicted by the high classification performance even for 2D and 3D, t-SNE manages to clearly capture inherent class structure associations. Lastly, the Laplacian Eigenmap, Figure 3(e) and 3(f), creates globally sparse, yet locally smooth representations. As captured by the figures, the distinctly triangular form in 2D is revealed as a projected aspect of a more complex, yet smoothly connected 3D geometric structure. As evident by upper “ridge” of malignant (red) lesion points and broad cystic (yellow) “fin” on the left, the Laplacian Eigenmap also manages to capture inherent class associations.

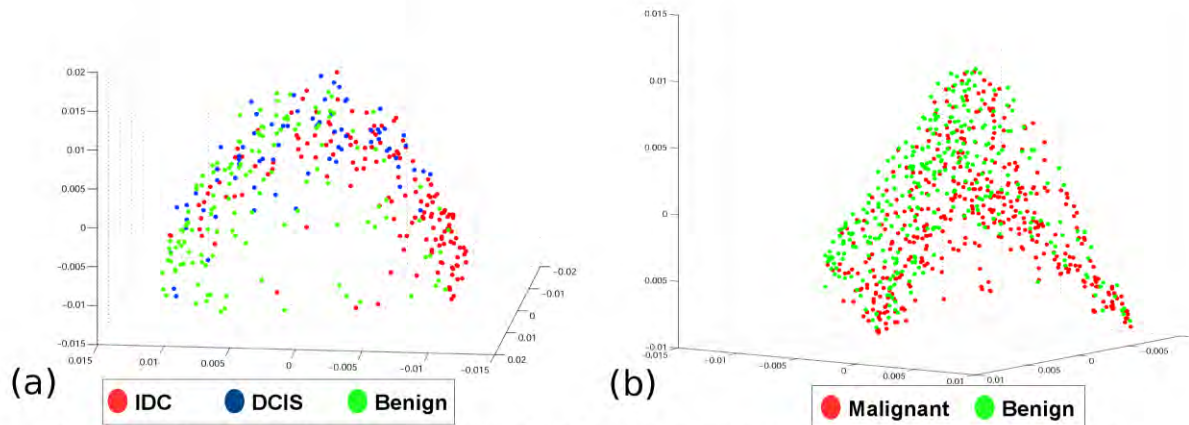


Figure 4. 3D visualization of the unsupervised local structure preserving non-linear dimension reduction representation using Laplacian Eigenmaps on breast lesion feature data. (a) 3D visualization of the entire originally 31 dimensional DCE-MRI feature data, green data points signify benign lesions, red: malignant-IDC, and blue: malignant-DCIS. (b) 3D visualization of the entire originally 40 dimensional FFDM feature data, green points for benign and red for malignant lesions.

The FFDM and DCE-MRI visual representations are noisier than the US due to the smaller sample size. A few examples are provided in Figure 4(a,b). The MRI dataset clearly exhibits a sparse arc-like geometric structure using the Laplacian Eigenmap. This structure seemingly separates the bulk of benign (green) lesions from the IDC (red) while dispersing the DCIS (blue) cases in between.

V. Discussion

V.A Dimension Reduction in CADx

Three major conclusions can be made regarding the use of DR techniques in breast CADx from this study. First, and most importantly, information critical for the classification of breast mass lesions contained within the original high-dimensional CADx feature vectors is not destroyed by applying the unsupervised, non-linear DR and representation techniques of t-SNE and Laplacian Eigenmaps. This observation is strongly supported by the robustness of the classification performance across the three different imaging modalities, US, DCE-MRI, and FFDM.

Second, according to the statistical re-sampling validation methods, the DR-based

classification performance characteristics appear to potentially rival or in some cases exceed that of traditional feature-selection based techniques. Additionally, both the linear PCA and non-linear t-SNE and Laplacian Eigenmap methods often generated “tighter” 95% empirical bootstrap intervals, implying reduced variance in classifier output, as compared to the feature selection based approaches, especially ARD, see Figure (4). For instance, in the large US dataset, the performance for 13 ARD selected features was $AUC_{0.632+} = 0.88$ with 95% empirical bootstrap interval $[0.787;0.895]$ and for 4 LSW selected features was $AUC_{0.632+} = 0.87$ with interval $[0.817;0.906]$ compared to 4D t-SNE mapping (from the original 81D feature space) giving $AUC_{0.632+} = 0.90$ with interval $[0.847;0.919]$. These findings imply that the generally non-linear manifold, on which US feature data exists, embedded in four dimensional Euclidean space can adequately represent the critical information for classification. These results build evidence for some potential benefits of employing the information-preserving, DR techniques in place of explicit feature selection, including the avoidance of the “curse of dimensionality”.

Third, the non-linear DR techniques generated visually-rich embedded mappings with a geometric structure that often presented sparse separation between class categories, as demonstrated in Figure 3(b): malignant, benign, cyst, and Figure 4(a): benign, DCIS, IDC. The natural class associations visible in the mappings are not totally unexpected since, as explored above, the classification performance results clearly demonstrate the reduced mapping’s capacity to retain sufficient information for class-discrimination. The large sample number of the US dataset provided the most vivid visualizations, highlighting both the geometric forms and sparse quality of the non-linear embeddings. Although PCA retained high supervised classification performance, unlike the non-linear Laplacian Eigenmaps and t-SNE embeddings, Figures 3(d),3(f), PCA is not capable of adequately representing the data’s inherent local structural properties, Figure 3(b), leading to less informative visualizations. Yet, the two non-linear methods offer distinct perspectives on the data

structures. The Laplacian Eigenmap appears to perhaps frame the lesions in a more globally smooth context as evidenced by the gradual transitions between distant regions of the geometric form, whereas t-SNE creates many distinct jagged “islands” of clustered lesion points. These emergent characteristics reflect the theoretically motivated principles driving the respective non-linear DR algorithms.

V.B Reduction Method Parameters

We briefly explored the impact of the parameter selection towards performance and visual appearance. To our knowledge there is no principled way to optimally select a parameter configuration, thus we simply choose parameters that gave reasonable mappings as discernable in the 2D/3D representations. This is a problem in general for many unsupervised techniques. In fact, as t-SNE creators noted², the method was primarily considered for visualization purposes and not explicitly for DR beyond 3 dimensions. Performance of t-SNE is not well understood for the general purpose of DR and subsequent classification. Future work may be of interest to discover procedures for identifying “optimal” or “near-optimal” subsets of parameters for CADx or similar machine learning purposes.

V.C. Classifiers and Feature Selection

In considering classifier design, one desires to be “as simple as possible, but no simpler,” meaning the most robust scheme in terms of both performance and stability (low variability in performance between different samples from the same underlying distribution), all while attempting to constrain the number of parameters, namely the input space dimension. Additionally, simpler models facilitate future repeatability with new contexts and datasets. The degree to which such pursuits are successful is dependent upon the interplay of the three main aspects affecting the

performances of the classifiers including: sample size, data complexity, and model complexity/regularization. Naturally included within the scope of the model complexity/regularization is the choice of inputs to the classifier, whether in the form of DR mappings or a set of selected features, as this also critically influences ultimate classification capability. Ideally, any classifier's aim is to synthesize the information available from the input space in a complete and unbiased fashion towards accomplishing the decision task. In general, classification of new input based on finite training dataset is an "ill-posed" problem, and regardless of the sophistication of regularization employed, instability may persist.¹⁵ For these reasons both the LDA and MCMC-BANN were investigated. By spanning over three different imaging modalities of varying data set size, using two different classifiers, and employing three different feature space approaches, all three of these key concepts (sample size, sample complexity, and model complexity) were touched upon in the course of this investigation.

For the relatively large US dataset, with 1126 unique lesions making up 2956 lesion images, some of the relative strengths associated with the more general, non-linear MCMC-BANN were particularly apparent. Specifically, the MCMC-BANN, when paired with either the DR techniques or LSW-FS was able to achieve high $AUC_{0.632+}$ performance, even at low input space dimensions, as seen in Figure 1(a). This is in part due to the MCMC-BANN ability to generalize to any target distribution, yet remain relatively well regularized, thereby avoiding "over-fitting" and severe underperformance on testing data. Yet, critically, when relying on explicit feature selection, across all input space dimension sizes for the FFDM and MRI data, and when fewer than 9 features were selected for the US data, the MCMC-BANN's success was contingent upon the use of LSW-FS over ARD-FS. The MCMC-BANN severely underperformed when coupled with the ARD-FS, especially when limited to picking only a few features. The smaller $AUC_{0.632+}$ and higher bootstrap variability (most dramatically evident for the lower input space dimensions), reveals limitations in ARD-FS

ability to consistently identify smaller sub-sets of features capable of robustly contributing to the classification task. This limitation may be in part due to ARD's capacity for discovering non-linear associations, which may vary highly between different bootstrapped sub-samples, as well as its less direct approach (compared to LSW) in feature determination.

Turning to LDA, while not best suited to model the non-linear DR mappings, the robustness and stability of LDA shines when joined with LSW-FS for classification purposes. LDA is, in a sense, naturally regularized by its linear nature and thus automatically avoids severe over-fitting situations. Often, the relative advantage of a more complex classifier, such as MCMC-BANN, over LDA, may begin to erode as sample size decreases, even if the underlying distribution is not completely linear in nature. These phenomena are apparent for the much smaller FFDM (245 unique cases, on 735 images) and DCE-MRI (356 unique lesions/images) datasets, as the less sophisticated LDA often produced the highest $AUC_{0.632+}$ values. The LDA classifier showed the greatest strength with the MRI data, nearly matching the LSW-MCMC-BANN and similarly for the DR approaches.

Furthermore, in examining Figure 2 again, among points falling within desirable performance specifications (upper left-hand corner: high classification performance/lower expected variability), it is reasonable to favor configurations which require the lowest input space dimensionality, as discussed previously (either the number selected features or target embedded mapping dimensions). A potential advantage of DR is that it may reduce the amount of necessary parameters (not including the unsupervised transformation characterized by the data itself) required to form a satisfactory data representation suitable for robust classification. In fact, most motivation for performing DR is lost if the target dimension is not considerably lower than the original high dimensional space. This is because such mapped representations become less efficient compared to simply making use of the original feature space or selected sub-space as dimensions are added.

Thus, within the framework of these criteria, in reviewing the results from the three modalities on whole, one may postulate, that as an overall strategy, 4D t-SNE appears likely to produce competitive classification performance when used as input into a non-linear classifier such as the MCMC-BANN. Such classification performance coupled with the intriguing 2D and 3D visualizations of the overall data structure may evoke attractive research potential.

In practice, it should be noted that, with the sole intention of maximizing classification performance based on finite sample training data, there may be no clear advantage for use of DR techniques over traditional feature selection. Although, again, due to the “curse of dimensionality,” as the input space dimension for classification becomes higher in dimension, eventually cross-validation based-performance will stagnant or even begin to regress lower. This occurs as the dataset sample size is not sufficient to adequately isolate a unique classifier solution (as many, potentially infinite, become possible) and marginal, if not none at all, new information is gained by the additional dimensions. Thus, for these reasons and in order to compare each dataset on common ground, the tests were limited to 1D-13D.

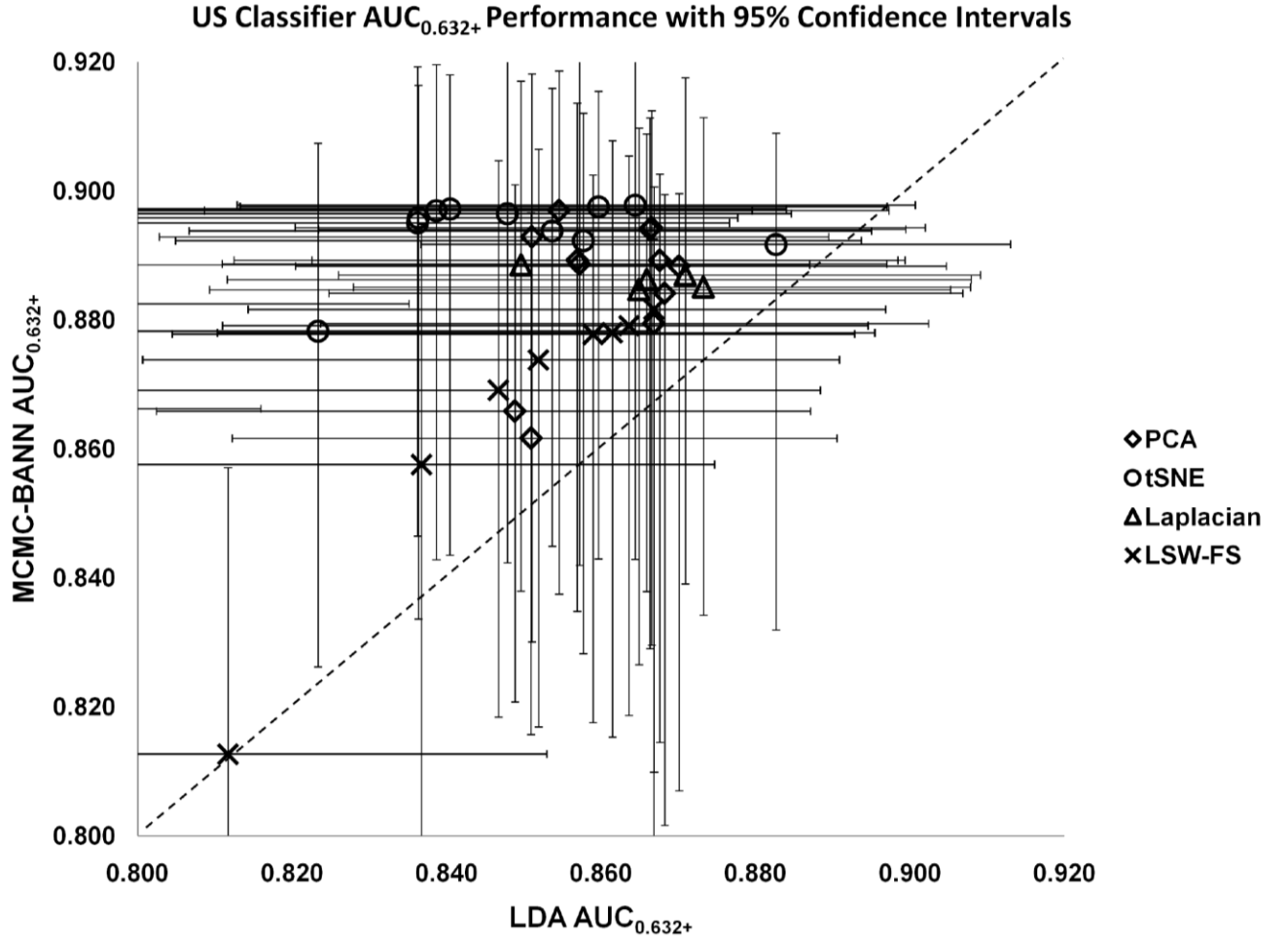


Figure 5. The 0.632+ bootstrapped area under the ROC curve is shown for MCMC-BANN (vertical axis) versus LDA (horizontal axis) with 95% empirical bootstrap confidence intervals included, for the originally 81 dimensional US feature dataset dimension reduced input or with LSW selected features.

VI. Conclusion

The ability to capture high-dimensional data structure in a human interpretable low-dimensional representation is a powerful research tool. The above findings strongly suggest the relevance of non-linear DR and representation techniques to future CADx research. DR cannot be expected to replace the benefits of feature selection based approaches in many cases. Yet, these techniques, in addition to competitive classification performance, do offer complementary

information and a fresh perspective on interpreting the overall structure of the feature data. Of interest to future studies is to further investigate the origin, meaning, and physical interpretation of the discovered structures present in the CADx lesion data as revealed by these non-linear, local-geometry preserving representations. Such rich data structure representations may offer novel insights and useful understandings of clinical CADx image data.

VII. Acknowledgments

This work is partially supported by US DoD Grant No. W81XWH-08-1-0731, from the US Army Medical Research and Material Command, NIH grants P50-CA125138, and DOE grant DE-FG02-08ER6478. The authors would like to gratefully acknowledge Lorenzo Pesce, Richard Zur, Jun Zhang, and Partha Niyogi for their thoughtful discussion and insightful suggestions. Additionally, we thank Weijie Chen contributing breast MRI feature data. The authors are grateful to Geoffrey Hinton and Laurens van der Maaten for freely distributing their algorithm code as well as the very handy dimension reduction MATLAB toolbox. We would also like to gratefully acknowledge the SIRAF shared computing resource, supported in parts by grants NIH S10 RR021039 and P30 CA14599, and its excellent administrator, Chun-Wai Chan. And lastly we thank the reviewers for their useful suggestions.

VIII. Appendix

VIII. A. Laplacian Eigenmaps Algorithm Outline

Beginning with k input points, x_1, \dots, x_k , in R^l :

Step 1: *Construct the Adjacency Graph*: Generate a graph with edges connecting nodes i and j

if x_i and x_j are “close.” Closeness is defined by the nodes included in the N nearest neighbors. This relation is naturally symmetric between points i and j . The parameter N must be selected.

Step 2: Choosing Weight: The “heat kernel” is used to assign weights to edge connected nodes i

and j : $W_{ij} = \exp(-||x_i - x_j||^2/t)$. Otherwise use $W_{ij} = 0$ for unconnected vertices. See Belkin and Niyogi for kernel justification¹. The parameter t is user defined. If t is set very high, or approximately, $t = \infty$, the edge connected node weights are essentially $W_{ij} = 1$, this option can be used to avoid parameter selection.

Step 3: Computing Eigenmaps: Assuming a connected graph generated in step 1, G , solve for

the following eigenvector and eigenvalues: $Lf = \lambda Df$, where D is the diagonal weight

matrix, defined by summing over the rows of W . $D_{ii} = \sum_j W_{ij}$, and L is the Laplacian matrix defined as: $L = D - W$. Symmetric and positive semi-definite, conceptually the Laplacian matrix acts as an operator on functions defined by graph G 's vertices.

Solving the equation, let f_0, \dots, f_{k-1} be the eigenvectors, arranged in accordance to their eigenvalues: $0 = \lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_k$. $Lf_0 = \lambda_0 Df_0 \dots Lf_{k-1} = \lambda_{k-1} Df_{k-1}$.

Finally, the k input data points in R^l are embedded in m -dimensional Euclidean space using the m eigenvectors after the zero eigen-valued f_0 , $x_i \rightarrow (f_1(i), \dots, f_m(i))$.

VIII. B. t-SNE Algorithm Outline

Beginning with k input points, $\{x_1, \dots, x_k\}$ in \mathbb{R}^l , set Perplexity parameter, $Perp$, number of iterations T , learning rate η , and momentum $\alpha(t)$.

Step 1. Compute Similarities: Compute pairwise $p_{j|i}$ probabilities using the σ_i found with perplexity $Perp$, and use symmetrized conditional probability distributions

$$p_{ij} = (p_{j|i} + p_{i|j})/2k$$

Step 2. Initialize Solution Sample: Sample from $N(0, 10^{-4}I^m)$ for initial points $\{y_1, \dots, y_k\}$

Step 3. Execute T Update Iterations on Y : Compute low-dimension similarities q_{ij} using eq.

(4)

and gradient using eq(5). Update Y using $Y^{(t)} = Y^{(t-1)} + \eta \frac{\partial C}{\partial y_i} + \alpha(t)(Y^{(t-1)} - Y^{(t-2)})$

Output: Low-dimension mapping $\{y_1, \dots, y_k\}$ in \mathbb{R}^m

IX. References

- ¹ M. Belkin, and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," Neural Comput. **15**, 1373--1396 (2002).
- ² L. van der Maaten, and G. Hinton, "Visualizing Data Using T-SNE," J. Mach. Learn. Res. **9**, 2605, 2579 (2008).
- ³ M.L. Giger, H. Chan, and J. Boone, "Anniversary Paper: History and Status of CAD and Quantitative Image Analysis: The Role of Medical Physics and AAPM," Med. Phys. **35**, 5799-5820 (2008).
- ⁴ Z. Huo, M.L. Giger, C.J. Vyborny, D.E. Wolverton, R.A. Schmidt, and K. Doi, "Automated Computerized Classification of Malignant and Benign Masses on Digitized Mammograms," Acad. Radiol. **5**, 155-168 (1998).
- ⁵ M. Kupinski, and M. Giger, "Automated Seeded Lesion Segmentation on Digital Mammograms," IEEE Trans. Med. Imaging **17**, 510-517 (1998).
- ⁶ Z. Huo, M.L. Giger, C.J. Vyborny, U. Bick, P. Lu, D.E. Wolverton, and R.A. Schmidt, "Analysis of Spiculation in the Computerized Classification of Mammographic Masses," Med. Phys. **22**, 1569-1579 (1995).
- ⁷ K. Drukker, M.L. Giger, K. Horsch, M.A. Kupinski, C.J. Vyborny, and E.B. Mendelson, "Computerized Lesion Detection on Breast Ultrasound," Med. Phys. **29**, 1438-1446 (2002).
- ⁸ W. Chen, M.L. Giger, U. Bick, and G.M. Newstead, "Automatic Identification and Classification of Characteristic Kinetic Curves of Breast Lesions on DCE-MRI," Med. Phys. **33**, 2878-2887 (2006).
- ⁹ W. Chen, "Computerized Interpretation of Breast MRI: Investigation of Enhancement-Variance Dynamics," Med. Phys. **31**, 1076 (2004).

- ¹⁰ K. Drukker, M.L. Giger, C.J. Vyborny, and E.B. Mendelson, "Computerized Detection and Classification of Cancer on Breast Ultrasound," *Acad. Radiol.* **11**, 526-535 (2004).
- ¹¹ M. Giger, "Computer-Aided Diagnosis of Breast Lesions in Medical Images," *Comput. Sci. Eng.* **2**, 39-45 (2000).
- ¹² G.D. Tourassi, B. Harrawood, S. Singh, J.Y. Lo, and C.E. Floyd, "Evaluation of Information-Theoretic Similarity Measures for Content-Based Retrieval and Detection of Masses in Mammograms," *Med. Phys.* **34**, 140-150 (2007).
- ¹³ Y. Yuan, M.L. Giger, H. Li, K. Suzuki, and C. Sennett, "A Dual-Stage Method for Lesion Segmentation on Digital Mammograms," *Med. Phys.* **34**, 4180-4193 (2007).
- ¹⁴ K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. (Academic Press, Boston, 1990).
- ¹⁵ C.M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006).
- ¹⁶ S. Geman, E. Bienenstock, and R. Doursat, "Neural Networks and the Bias/Variance Dilemma," *Neural Comput.* **4**, 1-58 (1992).
- ¹⁷ B. Sahiner, H. Chan, N. Petrick, R.F. Wagner, and L. Hadjiiski, "Feature Selection and Classifier Performance in Computer-Aided Diagnosis: The Effect of Finite Sample Size," *Med. Phys.* **27**, 1509-1522 (2000).
- ¹⁸ M.A. Kupinski, and M.L. Giger, "Feature Selection with Limited Datasets," *Med. Phys.* **26**, 2176-2182 (1999).
- ¹⁹ W. Chen, R.M. Zur, and M.L. Giger, "Joint feature selection and classification using a Bayesian neural network with automatic relevance determination priors: potential use in CAD of medical imaging" in *Medical Imaging 2007: Computer-Aided Diagnosis*, edited by M. Giger and N. Karssemeijer (2007), vol. 6514 of *Proc. SPIE*, pp. 65141G-10.
- ²⁰ M.A. Anastasio, H. Yoshida, R. Nagel, R.M. Nishikawa, and K. Doi, "A Genetic Algorithm-Based Method for Optimizing the Performance of a Computer-Aided Diagnosis Scheme for Detection of Clustered Microcalcifications in Mammograms," *Med. Phys.* **25**, 1613-1620 (1998).
- ²¹ G.D. Tourassi, E.D. Frederick, M.K. Markey, and J. Floyd, "Application of the Mutual Information Criterion for Feature Selection in Computer-Aided Diagnosis," *Med. Phys.* **28**, 2394-2402 (2001).
- ²² Y. Wang, D.J. Miller, and R. Clarke, "Approaches to Working in High-Dimensional Data Spaces: Gene Expression Microarrays," *Br. J. Cancer* **98**, 1023-1028 (2008).
- ²³ H. Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components," *J. Educ. Psychol.* **24**, 498-520 (1933).
- ²⁴ M. Kirby, *Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns* (John Wiley & Sons, Inc., New York, 2000).
- ²⁵ K. Drukker, N.P. Gruszaszkas, and M.L. Giger, "Principal component analysis, classifier complexity, and robustness of sonographic breast lesion classification", *Medical Imaging 2009: Computer-Aided Diagnosis*, edited by M. Giger and N. Karssemeijer (2009), vol. 7260, *Proc. in SPIE*, pp. 72602B-6.
- ²⁶ C. Varini, A. Degenhard, and T.W. Nattkemper, "Visual Exploratory Analysis of DCE-MRI Data in Breast Cancer by Dimensional Data Reduction: A Comparative Study," *Biomed. Signal Process. Control* **1**, 56-63 (2006).
- ²⁷ A. Madabhushi, P. Yang, M. Rosen, and S. Weinstein, "Distinguishing Lesions from Posterior Acoustic Shadowing in Breast Ultrasound Via Non-Linear Dimensionality Reduction," *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **1**, 3070-3073 (2006).
- ²⁸ M.K. Markey, J.Y. Lo, G.D. Tourassi, and C.E. Floyd, "Self-Organizing Map for Cluster Analysis of a Breast Cancer

Database,” *Artif. Intell. Med.* **27**, 113-127 (2003).

²⁹ K. Drukker, K. Horsch, and M.L. Giger, “Multimodality Computerized Diagnosis of Breast Lesions Using Mammography and Sonography,” *Acad. Radiol.* **12**, 970-979 (2005).

³⁰ H. Chan, D. Wei, M.A. Helvie, B. Sahiner, D.D. Adler, M.M. Goodsitt, and N. Petrick, “Computer-Aided Classification of Mammographic Masses and Normal Tissue: Linear Discriminant Analysis in Texture Feature Space,” *Phys. Med. Biol.* **40**, 857-876 (1995).

³¹ B. Sahiner, H. Chan, and L. Hadjiiski, “Classifier Performance Prediction for Computer-Aided Diagnosis Using a Limited Dataset,” *Med. Phys.* **35**, 1559 (2008).

³² R.M. Neal, *Bayesian Learning for Neural Networks* (Springer-Verlag New York, Inc., 1996).

³³ M. Kupinski, D. Edwards, M. Giger, and C. Metz, “Ideal Observer Approximation Using Bayesian Classification Neural Networks,” *IEEE Trans. Med. Imaging* **20**, 886-899 (2001).

³⁴ M.E. Tipping, in *Advanced Lectures on Machine Learning* (Springer, Berlin / Heidelberg, 2004), pp. 41-62.

³⁵ I. Nabney, *Netlab* (Springer, 2002).

³⁶ E. Levina, and B. Bickel, in *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA, 2005).

³⁷ M. Belkin, and P. Niyogi, “Towards a Theoretical Foundation for Laplacian-Based Manifold Methods,” *J. Comput. Syst. Sci.* **74**, 1289-1308 (2008).

³⁸ L. van der Maaten, “Matlab Toolbox for Dimensionality Reduction ” (2008).

³⁹ G. Hinton, and S. Roweis, in *Advances in Neural Information Processing Systems 15* (The MIT Press, Cambridge, 2003), pp. 833-840.

⁴⁰ L. van der Maaten, “T-SNE Files” (2008).

⁴¹ L.L. Pesce, and C.E. Metz, “Reliable and Computationally Efficient Maximum-Likelihood Estimation of “Proper” Binormal ROC Curves,” *Acad. Radiol.* **14**, 814-829 (2007).

⁴² C.E. Metz, “Basic Principles of ROC Analysis,” *Semin. Nucl. Med.* **8**, 283-298 (1978).

⁴³ J.A. Hanley, and B.J. McNeil, “The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve,” *Radiology* **143**, 29-36 (1982).

⁴⁴ B. Efron, and R. Tibshirani, “Improvements on Cross-Validation: The .632+ Bootstrap Method,” *J. Am. Stat. Assoc.* **92**, 548-560 (1997).

Appendix C

